Abolfazl Asudeh

Fall 2020

10/08/2020

InDeX Lab
Innovative
Data
eXploration
Laboratory

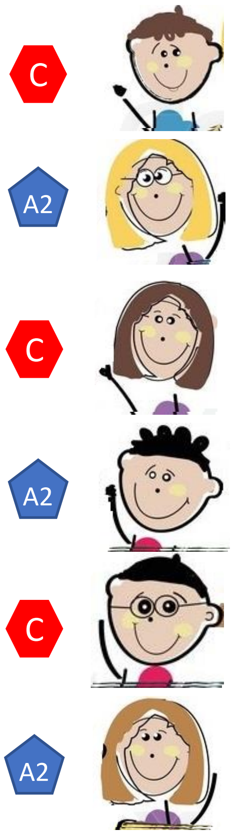
CS 594: In-process Interventions to Achieve Fairness

Ranking

Toy Example

A2 Ann Arbor

D Chicago



Suppose you own a real estate agency with two branches in Ann Arbor and Chicago.

You want to give bonus to

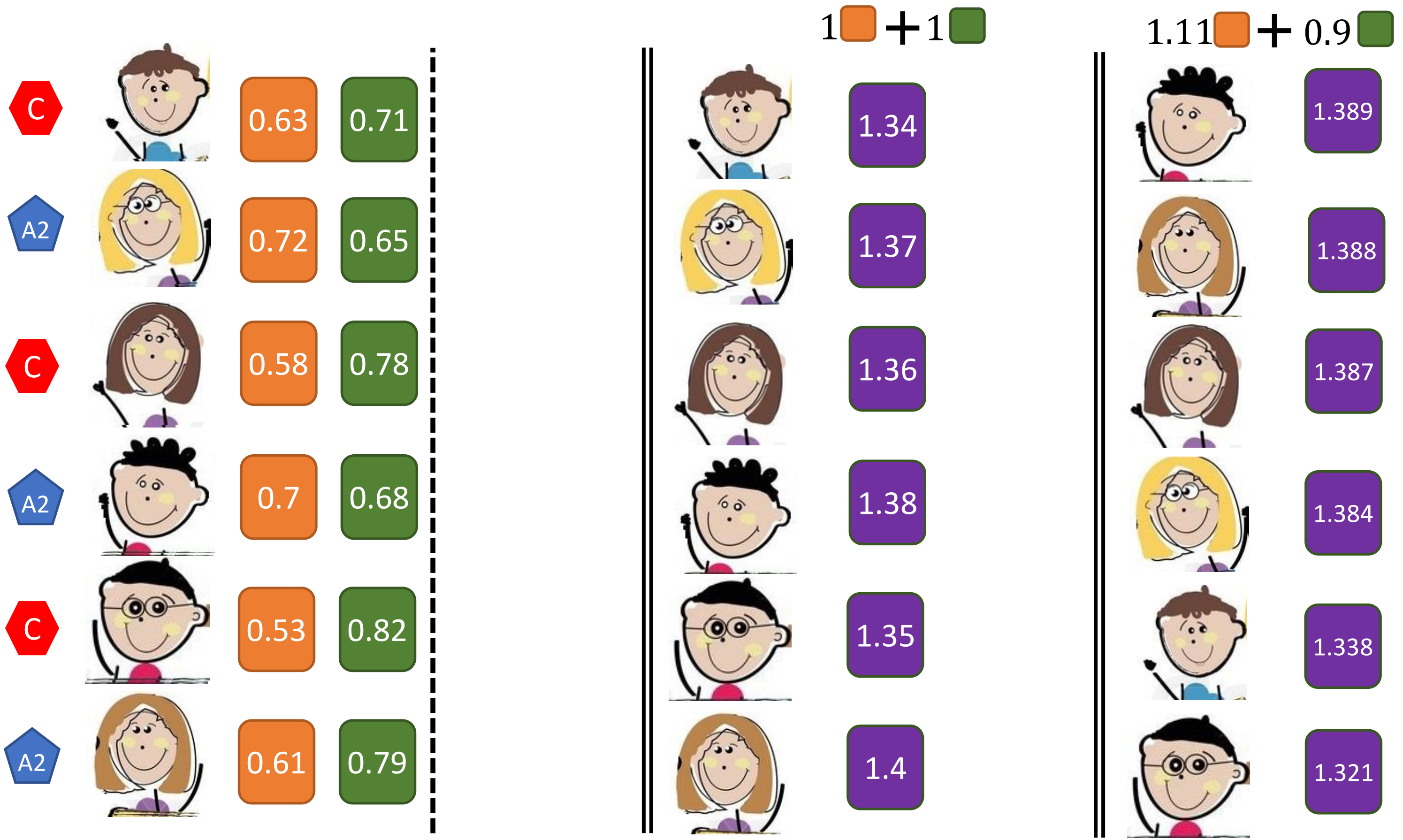
(1) Top-3 agents

To be fair, you want to make sure that each branch receives at least one promotion

Toy Example

Sale -- Normalized

Customer Satisfaction -- Normalized



$$\sum \theta_i x_i$$

$$\theta_1 \square + \theta_2 \square$$

Despite the potential impact of these weights, those are

chosen in an ad-hoc manner!

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

- "It is easy to see why the U.S. News rankings are so popular. A single score allows us to judge between entities"
- "Rankings depend on what weights we give to what variables"
- "This idea of using the rankings as a benchmark, college presidents setting a goal of 'We're going to rise in the *U.S. News* ranking' ..."

THE NEW YORKER

Rankings depend on what weight we give to what variables.

Illustration by SEYMOUR CHWAST

Designing Fair Ranking Schemes

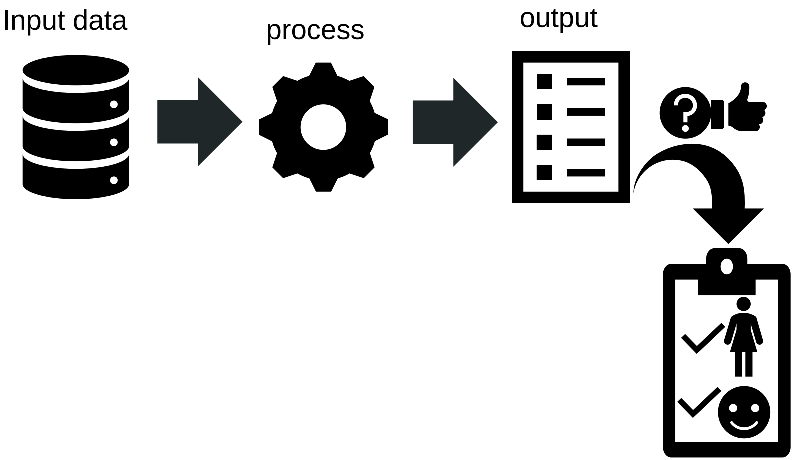
Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das

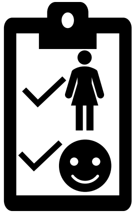
SIGMOD 2019

Fairness Model:

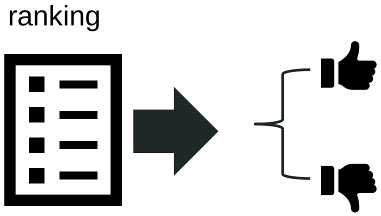
to support human values

- *Generate Fair outcomes*
- *Without Disparate Treatment*: ***explicit use of sensitive attributes*** to make decisions
    - not allowed in many jurisdictions





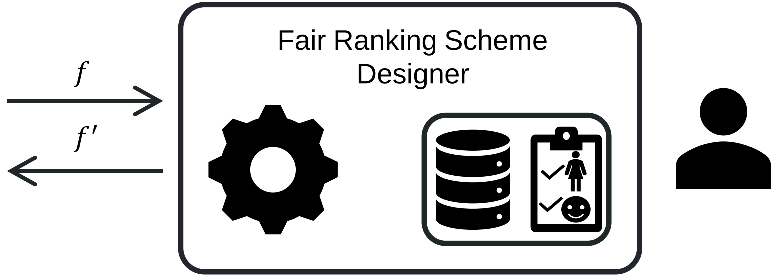
Fairness Generalization



High level idea

- *Offline*: Preprocess the data and generate some indices
  - OK not to be super fast

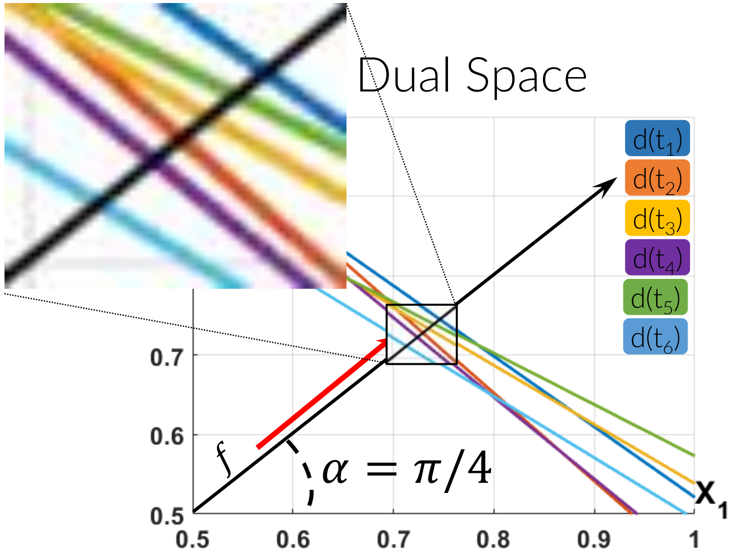
- *Online*: Answer user queries
  - Should be fast



2D Algorithm

Geometric interpretation

$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4



$d(t)$: $\sum t[i] \times x_i = 1$

2D:

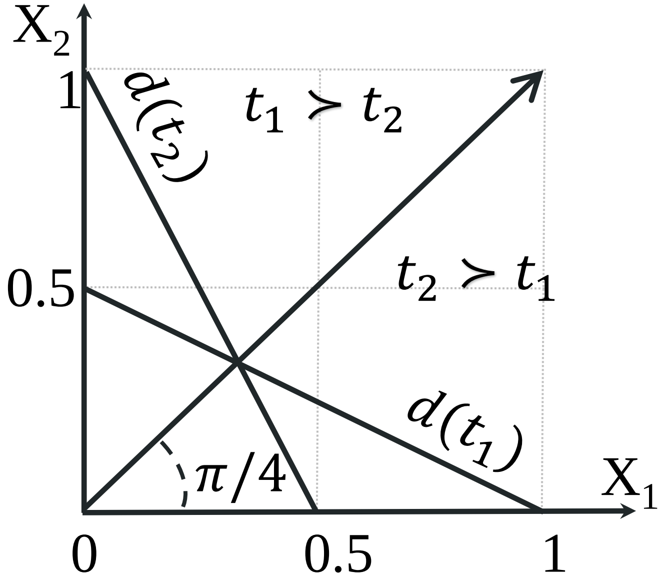
$d(t)$: $t[1]x_1 + t[2]x_2 = 1$

Ordering Exchange

- example

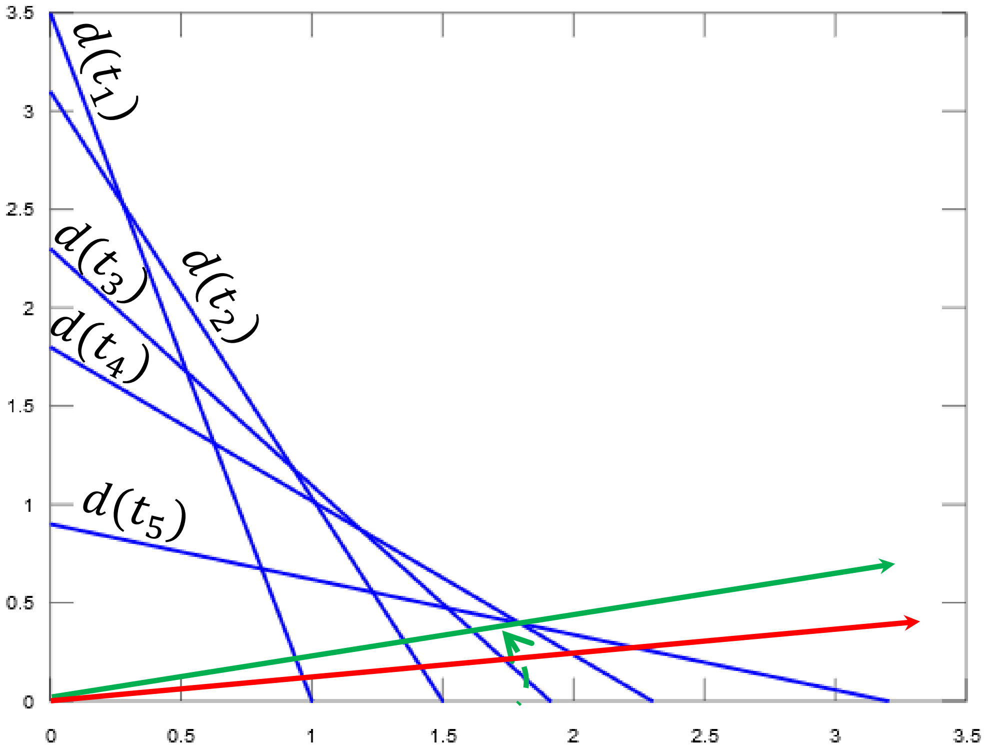
$t_1 < 1,2 >$

$t_2 < 2,1 >$



Ranking Regions

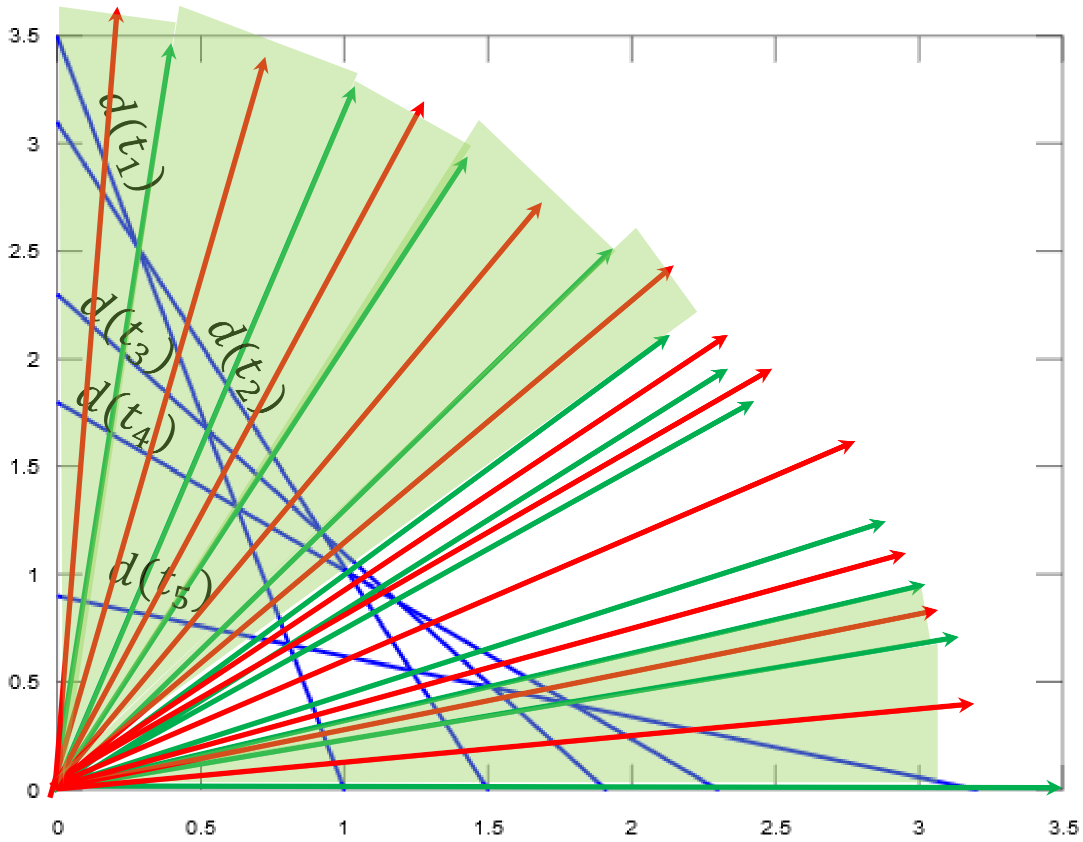
	x_1	x_2	location
t_1	3.5	1	A2
t_2	3.1	1.5	A2
t_3	2.3	1.91	C
t_4	1.8	2.3	C
t_5	0.9	3.2	A2



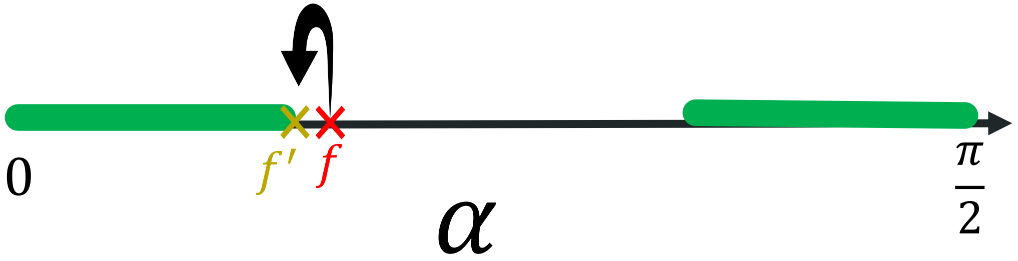
2D, offline:

	x_1	x_2	location
t_1	3.5	1	C
t_2	3.1	1.5	A2
t_3	2.3	1.91	C
t_4	1.8	2.3	A2
t_5	0.9	3.2	C

Fairness criterion:
at least one from each branch



2D: Online



- *Apply Binary Search!*

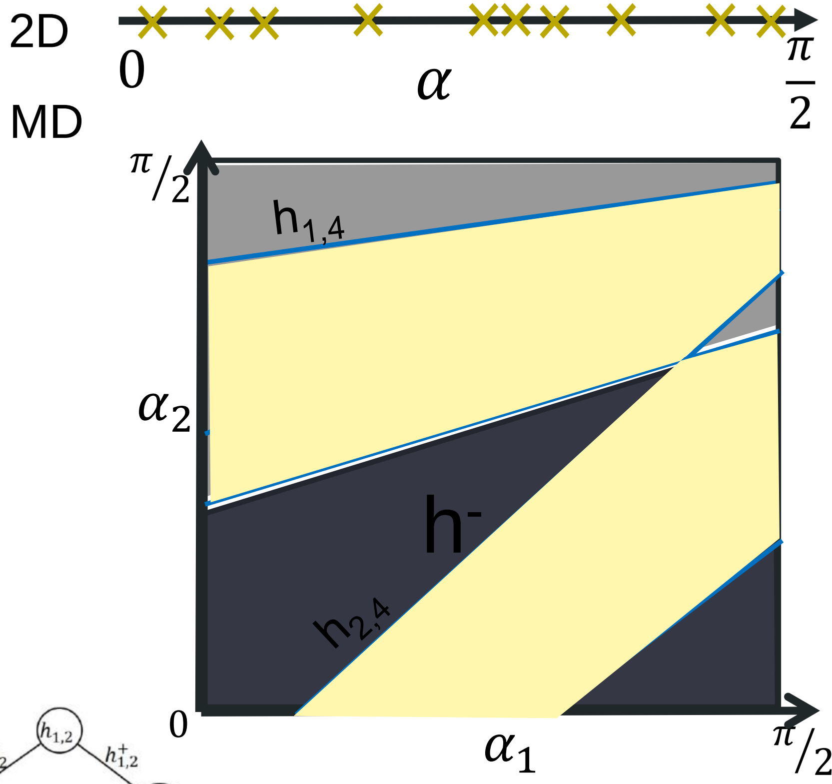
  fast: $O(\log n)$

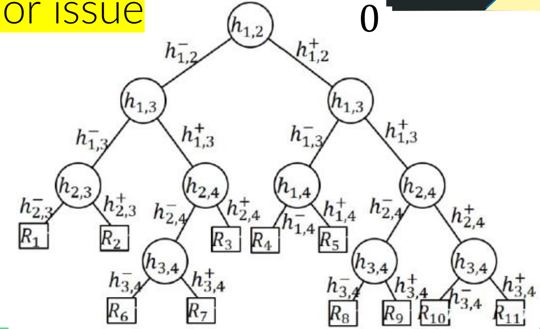
* n: number of items

MD Algorithm

MD (more than 2 attributes)

- 2D Extension:
  - Ordering Exchanges in MD
    - Half-spaces
  - Arrangement of Hyperplanes
    - $O(n^d)$.
    - d: number of attributes
  - Arrangement Tree
    - helps in practice
    - Still online processing is the major issue



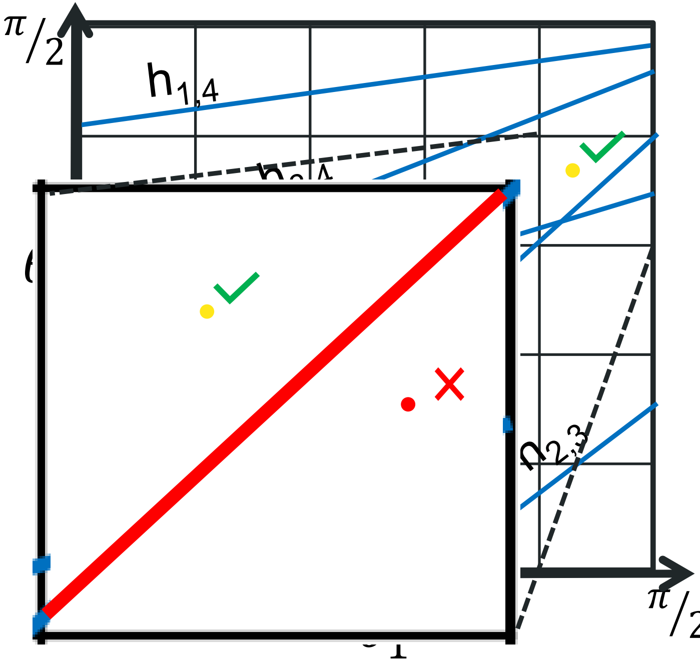


MD – approximation

- Trade-off accuracy with efficiency:

  Rather than “closest” fair function, return s/t within a constant additive approx. from the optimal “distance”

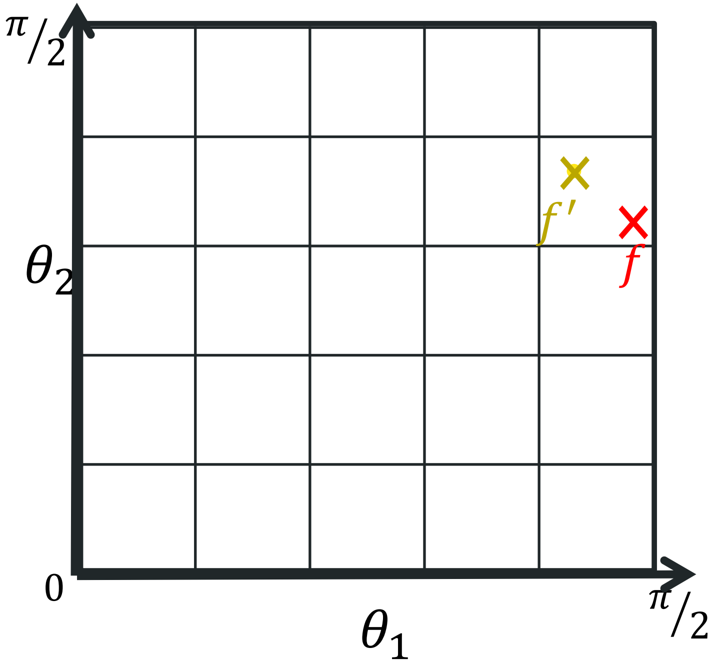
- At a high level:
  1. Partition the function space into equi-volume cells
     - The idea is to *assign a fair func. to each cell*
  2. Limit the arrangement to each cell and *stop* when found a fair function
  3. Assign the cells w/o a fair function to the closest discovered function



MD - Online

Simple:

1. Locate the cell to which the input function belong
2. Return the assigned function to the cell
   - fast: $O(\log N)$ – N is the number of cells



* N: number of cells in the partition

Scalability, On-the-fly query processing

- Uniform Item Sampling for scalability
  - Satisfactory functions over a uniform sample are *expected* to be satisfactory
- Uniform Function Sampling for on-the-fly processing
  - Negative result: cannot guarantee the discovery of a satisfactory function with any probability $p<1$
  - Still is expected to find “large” (stable) satisfactory regions

On obtaining stable rankings

Abolfazl Asudeh, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich

VLDB 2019

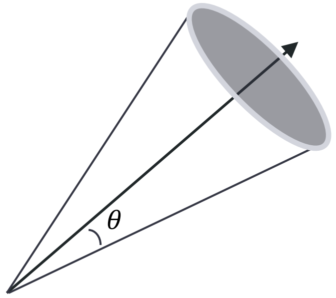
Stability: how robust the output is

- Small changes in weights change the output?
    - Decisions based on which are questionable (not fair)
    - Not Stable

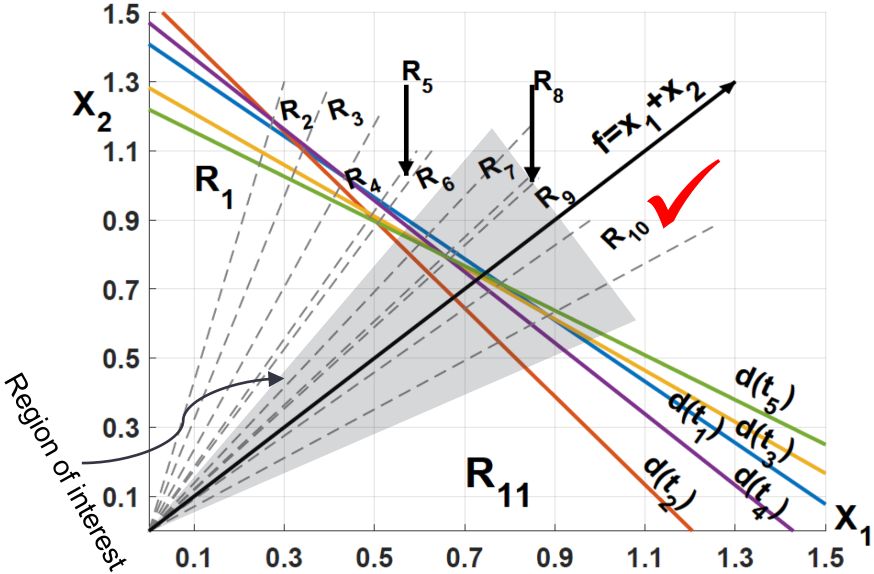
- Stability: The (volume) Ratio of functions that generate an output (ranking, top-k, or partial ranking)

Region of Interest

- The range of weights that are “acceptable” to the ranking designer
    - *A vector and angle distance: e.g.* at least 95% cosine similarity with a ref. vector



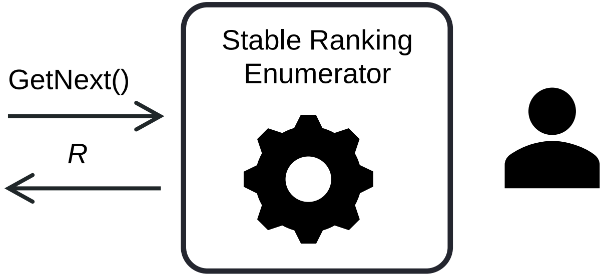
$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.83	0.65	1.48
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35



High level idea

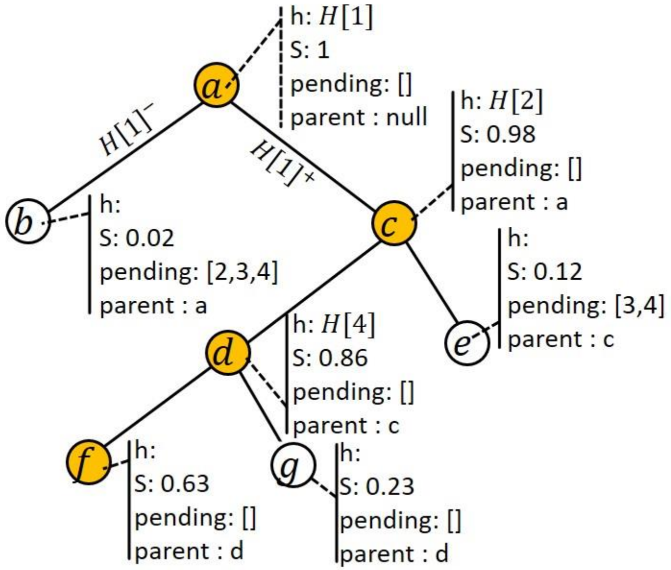
- *GetNext*: An iterative process that generate stable regions one after the other

- The user can keep enumerating stable rankings (or top-k), until he finds a satisfactory one



MD -- Threshold-based Algorithm

- Uses the arrangement tree
- In high-level:
    - Constructs the arrangement tree while only adds a postponing the process for the smaller regions

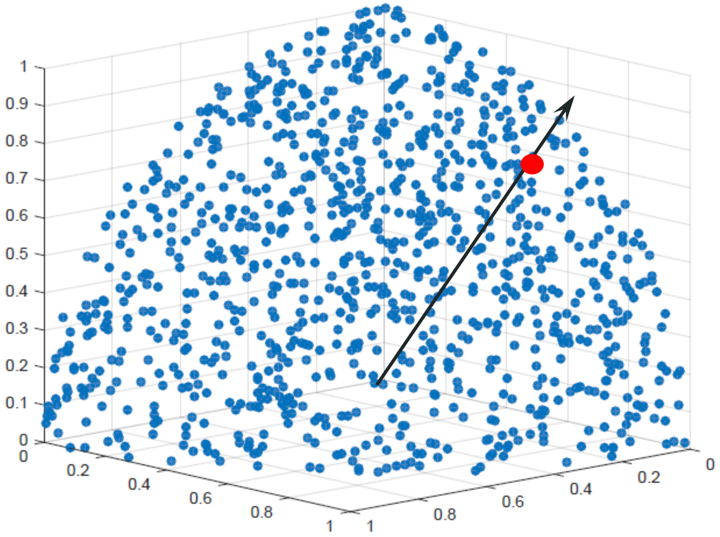


Randomized Get-Next

- A Monte-Carlo method that work based on repeated sampling and the central limit theorem

Unbiased sampling from the function space

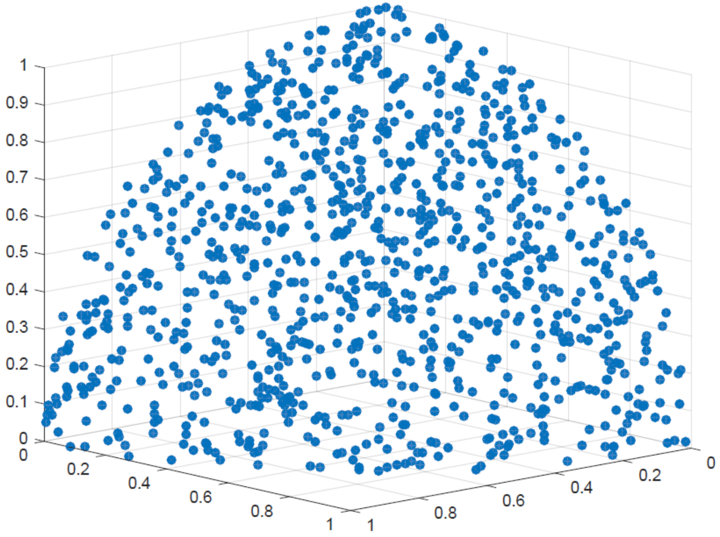
- 1-1 mapping b/w the functions (origin-starting rays) and the points on the *surface* of origin-centered unit *d-sphere* (hypersphere in $\mathbb{R}^d$)



* d: number of attributes

Unbiased sampling from the function space

- Sampling the weights Uniformly? ✗
- Sampling the weights using the Normal distribution ✓



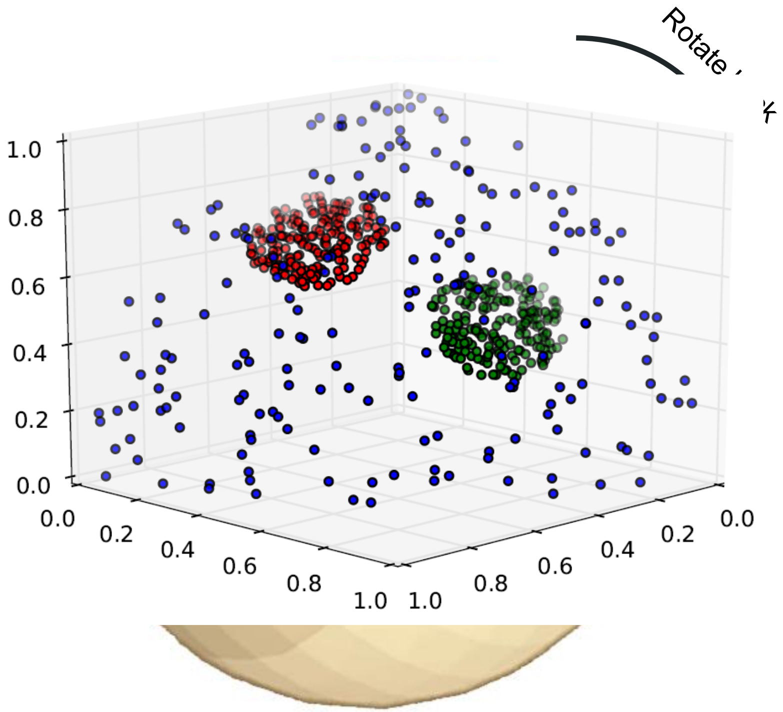
Asudeh, Abolfazl, and H. V. Jagadish. "Responsible Scoring Mechanisms Through Function Sampling." *arXiv preprint arXiv:1911.10073* (2019).

Sampling from a region of interest

- Each Riemannian Piece is a (d-1)D Sphere (ring in 3D)
- We know how to sample from its surface!: *Normal distribution*

- High-level:

1. Select each “ring” randomly, proportional to its area
2. Select a “point” from the surface of ring (using the Normal dist.)
3. Rotate the space back

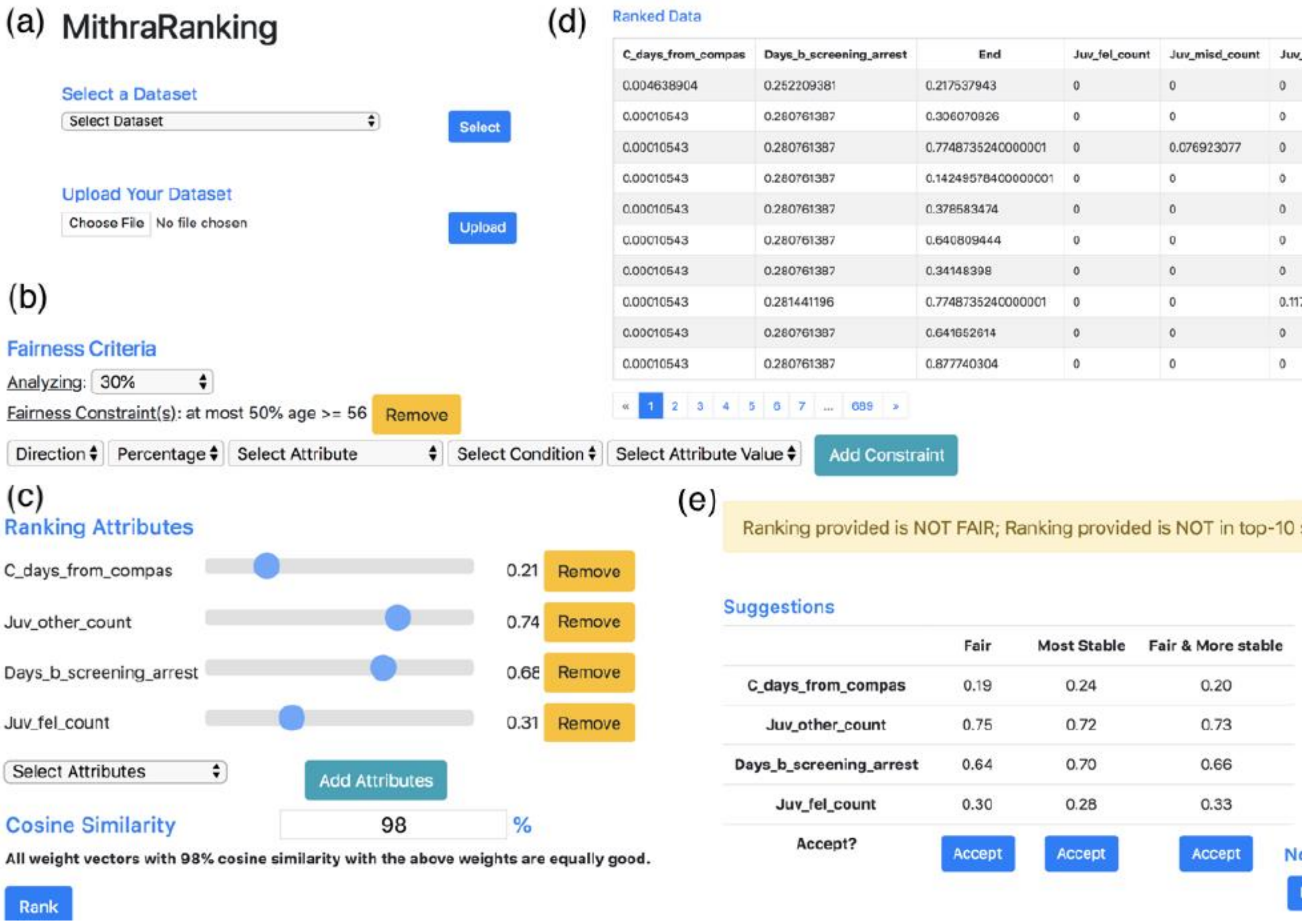


Randomized Get-Next

1. Take *N* unbiased sample functions from the region of interest
2. While keeping a hash of outputs, “count” the number of appearance for each output
3. return the output that appeared the most

    - estimate its stability & compute the confidence interval

MithraRanking



[*] Yifan Guan, Abolfazl Asudeh, Pranav Mayuram, H. V. Jagadish, Julia Stoyanovich, Gerome Miklau, and Gautam Das. Mithraranking: A system for responsible ranking design. SIGMOD 2019.

Nutritional Labels

Nutritional labels for interpretability

- Interpretability is an essential ingredient of successful machine-assisted decision-making.
- This motivates creating tools that show deficiencies, biases, and unfairness in score-based evaluation.

- Drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes:
    - a nutritional label is a set of automatically constructed visual widgets, each conveying standardized information about "fitness for use" of data or the evaluators

[*] Julia Stoyanovich, and Bill Howe. "Nutritional Labels for Data and Models." IEEE Data Eng. Bull. 42, no. 3 (2019): 13-23.

Ranking Facts: Nutritional Labels for Rankers

Ranking Facts

Recipe

Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

Attribute	Importance
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Ingredients

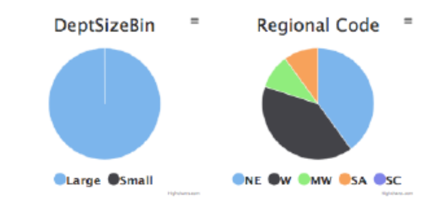
Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

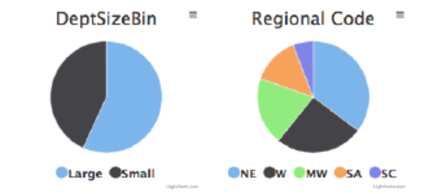
Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Diversity at top-10



Diversity overall



Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

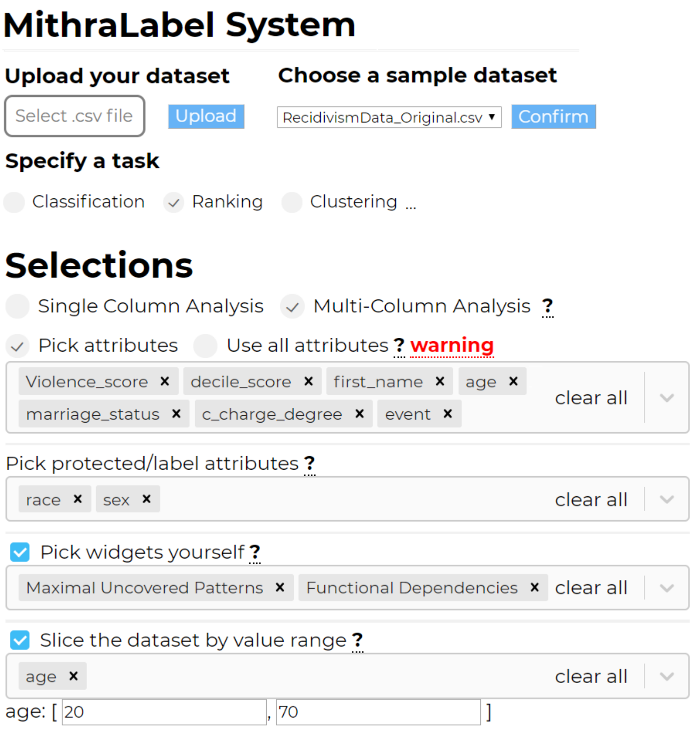
Fairness

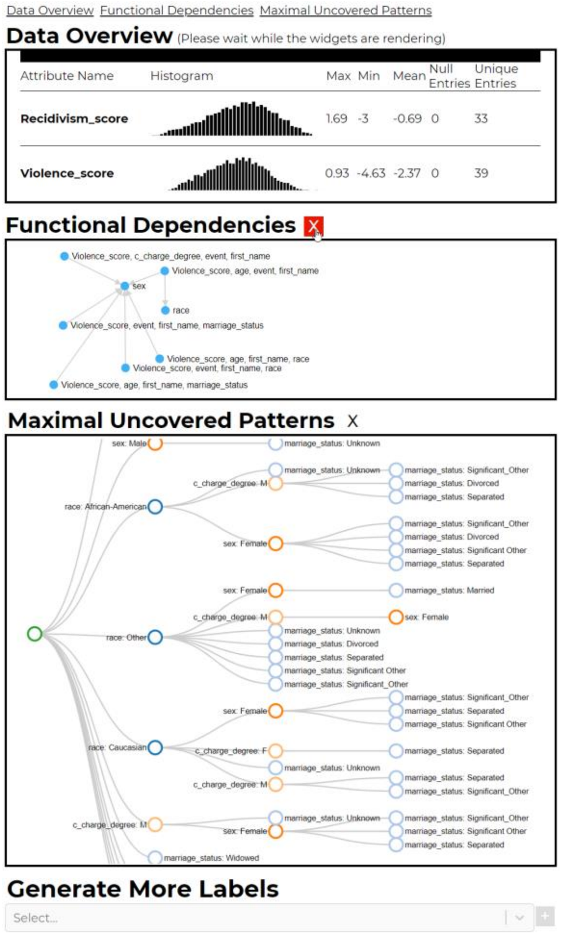
	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.98	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

FA*IR and difference in proportions (Proportion) are measured with respect to 26 highest-scoring items (the top-K). The top-K contains 100 items or one half of the input, whichever is smaller.

[*] Ke Yang, Julia Stoyanovich, A. Asudeh, Bill Howe, H. V. Jagadish, and G. Miklau.
A nutritional label for rankings. In SIGMOD 2018.

MithraLabel: Flexible Data set Nutritional Labels





[*] C. Sun, A. Asudeh, H. V. Jagadish, B. Howe, and J. Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In CIKM 2019